DOCUMENT RESUME

ED 286 902 TM 870 533

AUTHOR Webber, Larry; And Others
TITLE Estimating the Reliability of Dynamic Variables Requiring Rater Judgment: A Generalizability Paradigm.
SPONS AGENCY National Heart, Lung, and Blood Inst. (DHHS/NIH), Bethesda, MD.
PUB DATE 19 Nov 86
GRANT NRDC-A-HL15103
NOTE 23p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Memphis, TN, November 19-21, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; Cardiovascular System; *Error of Measurement; *Generalizability Theory; Health Education; Interrater Reliability; Mathematical Models; *Measurement Techniques; *Observation; *Reliability; Research Design; Research Problems
IDENTIFIERS Systolic Blood Pressure

ABSTRACT

Generalizability theory, which subsumes classical measurement theory as a special case, provides a general model for estimating the reliability of observational rating data by estimating the variance components of the measurement design. Research data from the "Heart Smart" health intervention program were analyzed as a heuristic tool. Systolic blood pressure readings for 17 children were taken by two nurses in one of six pairs. Differentiating the sources of error variance was carried out in what is called a "G" study, which used the GENOVA program to estimate the variance of measurement facets. As expected, the variance of individuals' blood pressure was large relative to pairs of nurse observers, sets of measurement, and occasions of measurement within the sets. A generalizability coefficient is the expected squared correlation between the actual, observed scores and the full universe of scores defined by the researchers by specifying measurement facets or protocols. It allows the calculation of different coefficients, given interest in making either norm-referenced or criterion-referenced decisions. While the "G" study evaluated data quality for different decisions, a "D" study (i.e., a study estimating how many conditions are needed to obtain a certain level of generalizability) was also conducted to pinpoint sources of measurement error, so that observations could be increased on a particular facet to reduce error. Many levels for each protocol facet are needed to estimate this properly. (LPG)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

12/18/86

ED286902

# ESTIMATING THE RELIABILITY OF DYNAMIC VARIABLES REQUIRING RATER JUDGMENT: A GENERALIZABILITY PARADIGM

Larry Webber Bruce Thompson Gerald Berenson

Louisiana State University Medical Center

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Thompson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN, November 19, 1986. This research was supported by funds from the National Heart, Lung and Blood Institute of the United States Public Health Service, National Research and Demonstration Center--Arteriosclerosis (NRDC-A) grant HL15103, "Heart Smart" School Intervention Program. Address reprint requests to Gerald S. Berenson, MD, Department of Cardiology, LSU Medical Center, 1542 Tulane Avenue, New Orleans, LA 70112-2822.

TM 870 533

ABSTRACT

The major benefits of generalizability theory, particularly as the theory can be applied to evaluate observational data, are enumerated. Actual research data from the "Heart Smart" health intervention program are employed for heuristic value to make the discussion concrete. It is suggested generalizability theory forces the researcher to make careful decisions about what measurement questions are being asked and about the population to which results are to be generalized. The theory is also valuable in that it allows the calculation of different coefficients given interest in making either norm-referenced or criterion-referenced decisions. Finally, the theory supports the analysis of "what if" questions that evaluate the value of variations in the measurement protocol and the economics of the protocol.

As Best (1981, p. 158) has noted, "From the earliest history of scientific activity, observation has been the prevailing method of inquiry." Indeed, Mouly (1978, p. 225) has suggested that "Research in the behavioral sciences is often concerned with phenomena whose status can only be estimated on the basis of [partially] subjective judgment" by observers. However, the validity of observational data can be compromised by a variety of factors, which have come to be recognized as the components of a catalog of threats to the validity of such data. Three major threats to the validity of observational data have been noted in the literature.

First, as Cates (1985, p. 224) notes, "Errors of leniency occur when observers tend to rate behaviors on evaluative scales higher than they should be rated because of a feeling that a lower rating might be 'a bit hard on' the subject." Second, as Borg and Gall (1983, p. 483) explain, "The so-called halo effect... is the tendency for the observer to form an early impression of the person being observed and to permit this impression to influence his ratings on all behaviors involving the given individual." Third, Cates (1985, p. 102) notes that "Errors of central tendency occur when an observer who is having difficulty rating a behavior... resolves the difficulty by assigning such behaviors to the central (or average) portion of the scale."

Researchers employing observational data also confront a dilemma regarding training. The failure to train observers to employ measurement protocols may result in intraindividually unstable ratings over subjects or in contradictory ratings by

different observers rating the same subject. However, as Mouly (1978, p. 225) has argued, "the training of observers sensitizes them in the direction of the investigator's biases." This sensitization may make observers aware of specific expectations, and "contamination... may occur if the observer is knowledgeable about the specifics of the study" (McMillan & Schumaker, 1984, p. 158).

Finally, threats to validity involving the subjects themselves, rather than the observers, can occur. The guniea pig effect occurs

> ...because of the subject's awareness of being tested. Although it does not necessarily follow that awareness leads to measurement errors or distortion, the probability of such errors increases as the subject's awareness of being measured increases. (Borg & Gall, 1983, p. 496)

For example, Mercatores and Craighead (1974) found that observation in classrooms tended to increase the frequency of teacher-pupil interaction. Some researchers attempt to avoid these problems by employing "non-reactive" or unobtrusive measurement methods. But these measures may not be available for certain variables in some studies. Furthermore, as Borg and Gall (1983, p. 499) note, "Reliability is also a problem with many nonreactive measures; many of these measures are essentially similar to a one-item test or to one question from a questionnaire."

Because "direct observation is time-consuming, and its cost

in money, as well as time, is usually considerable" (Hopkins, 1980, p. 54), and because "the validity and reliability of observation as a research technique depends critically on the competence of the observer" (Mouly, 1978, p. 224), some researchers go to considerable effort to establish that their observational data are meaningful. Yet Salvia and Mersel (1980) reviewed 153 observational studies with a high potential for bias and reported that only 22% reported adequate safeguards. Furthermore, as Rowley (1976, p. 51) notes, "It has been common to avoid the question of reliability altogether, or else to report a coefficient of observer agreement, knowing full well its inadequacy for that purpose." Dyer (1979, p. 124) succinctly pinpoints the inadequacy of using classical measurement theory to evaluate observational data:

> Although interobserver agreement is the most frequent technique used by researchers, this index does not reflect the consistency of the subject's behavior over time or the consistency with which the observation instrument distinguishes among individuals.

An alternative approach to using classical measurement theory to evaluate the measurement adequacy of observational data invokes the generalizability theory elaborated by Cronbach, Gleser, Nanda, and Rajaratnum (1972). Generalizability theory has been shown to be the most general measurement theory, i.e., subsumes classical measurement theory as a special case (Crocker & Algina, 1986).

The purpose of the present paper is to elaborate some of the

major benefits of generalizability theory, particularly as the theory can be applied to evaluate observational data. Actual research data from the "Heart Smart" health intervention program are employed for heuristic value to make the discussion concrete. More complete discussions of generalizability theory, and of the relationships between classical and generalizability measurement theories, are available in recent texts (Algina & Crocker, 1986; Brennan, 1983).

## Heuristic Data

In the classical measurement perspective, reliability is defined as the ratio of systematic variance to total variance in data. Cattell (1966), in his "data box," conceptualized three possible sources of variance in any data set: variations associated with which subjects are measured, variations associated with the variables on which measurements are taken, and variations associated with the use of one or more occasion of measurement.

Both classical measurement theory and the alternative measurement theory, i.e., generalizability theory, "assume that the phenomenon being studied remains constant over observations, i.e., is in a steady state" (Shavelson, Webb & Burstein, 1985, p. 72). Sax (1980, p. 261) concurs, noting that "Measures of stability are not appropriate if the trait being measured is itself unstable." This assumption can cause serious problems in evaluating the measurement properties of observation data, "because some behaviors may [by their nature] be more stable than others" (Dyer, 1979, p. 125). These problems may be particularly

characteristic of observational data. As Mouly (1978, p. 224) notes,

> In the dynamic type of situation where observation is presumably most appropriate, it is typically difficult to obtain relevant information sufficiently free from complicating co-occurences to give a clear picture of what is really involved.

An important advantage of generalizability theory is that it allows the researcher to estimate the amount of variance in data generated by each of the sources identified by Cattell (1966).

The heuristic data discussed here may be especially helpful in illustrating the process of estimating sources of variance within measurement protocols, because the variable measured, i.e., systolic blood pressure, is by its very nature somewhat dynamic and reactive. Furthermore, some observer judgment is required in the measurement process, as is typically the case with most observational data. Thus, special care must be taken with this type data to evaluate measurement characteristics.

The heuristic data are also useful because they force recognition that all variables are somewhat dynamic over occasions of measurement, and that therefore

> It is more accurate to talk about the reliability of <u>measurements</u> (data, scores, and observations) than the reliability of <u>tests</u> (questions, items, and other tasks). Tests cannot be stable or unstable, but observations can. Any reference to the "reliability of a test" should always be

> interpreted to mean the "reliability of measurements or observations [i.e., a particular set of data] derived from a test." (Sax, 1980, p. 261)

Rowley (1976, p. 53) concurs, noting that "It needs to be established that an instrument itself is neither reliable nor unreliable." Only specific data can be reliable. Researchers merely hope that findings regarding measurement qualities associated with a particular data set will generalize to future uses of the same measurement protocol with similar subjects.

The heuristic data involved real systolic blood pressure measurements taken from each of 17 individuals using mercury sphygmomanometers in good working order using a standardized protocol (Voors, Foster, Freuchs, Webber & Berenson, 1976). Each individual was randomly assigned to be measured by each member of one of six pairs of nurses. Thus, in analysis of variance terms, the measures of individuals were nested within the pairs of nurses. Each child was subjected to two sets of observations, i.e., one by each of the nurses in a given pair. Each set of measurements by a given nursse consisted of three observations, so observations were nested within sets. The design of the measurement is graphically presented in Figure 1.

INSERT FIGURE 1 ABOUT HERE.

## G-study Analyses

As Webb, Shavelson, Shea, and Morello (1981, p. 187) note, "Generalizability (G) theory evolved out of the recognition that

the concept of undifferentiated error in classical test theory provided too gross a characterization of the multiple sources of error in a measurement." The first step in differentiating these sources of error variance involves the estimation of the variance components associated with every facet of the measurement of the object being measured, i.e., what is called a "G" study.

Table 1 presents these estimates for the heuristic data. The results were generated by the GENOVA program documented by Crick and Brennan (1983). Shavelson and Webb (1981, p. 155) list other computer programs that have been specially developed to perform these analyses.

INSERT TABLE 1 ABOUT HERE.

The data in Table 1 can be evaluated on a preliminary basis to make some initial judgments of the quality of the data generated using the measurement protocol. Ideally one would want the variance component for the individials to be large relative to the facets of measurement used to collect the data. That is, it would be presumed that the subjects vary in their blood pressures, but it is hoped that the measurements are stable over pairs of nurse observers, sets of measurements, and occasions of measurements within the sets. In general these expectations are supported for these data.

However, a more straightforward interpretation is based on the calculation and interpretation of generalizability coefficients. Exploration of the nature of these coefficients gets at the essence of generalizability theory. As Cronbach et al. (1972, p. 15) explain:

> The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. The question of "reliability" thus resolves into a question of accuracy of generalization, or generalizability.

In contrasting the generalizability view of reliability with the classical measurement theory view, Shavelson, Webb and Burstein (1985, p. 62) note that

> G theory speaks of universe scores rather than true scores, acknowledging that there are different universes to which decision makers may generalize. Likewise, the theory speaks of generalizability coefficients rather than the reliability coefficient, realizing that the computed value of the coefficient may change as the definition of the universe changes.

A generalizability coefficient is the expected squared correlation between the actual, observed scores and the full universe of scores defined by the researcher by specifying measurement facets or protocols (Shavelson, Webb & Burstein, 1985, p. 63).

These concepts have great intuitive appeal for the researcher who is interested not only in establishing stability

of data, but who also wishes to establish generalizability to defined populations of people, variables, and occasions. As Rowley (1976, p. 55) notes:

> It would normally be the responsibility of the investigator to demonstrate that the observations he has obtained are indeed representative of the universe to which he claims to generalize. This should not be thought of as an imposition; in fact it ought to be regarded as essential if any generalization at all is to be made from the study.

This ought to be the primary focus of scientific inquiry.

Another positive feature of generalizability theory is that the theory recognizes that data may be employed in service of different types of decisions.

> For example, some interpretations may focus on individual differences (i.e. relative or comparative decisions), some may use the observed score as an estimate of a person's universe score (absolute decisions; cf. criterion-referenced interpretations). (Shavelson & Webb, 1981, p. 135)

Classical measurement theory typically emphasizes relative decisions; the coefficient that does so in generalizability analysis is the generalizability coefficient, i.e., the expected squared correlation coefficient between subjects' actual scores and their universe scores. For domain-or criterion-referenced decisions, an analogous generalizability coefficient, the phi coefficient, is of interest. For the data represented in Table 1, assuming primary interest in making statements about the quality

of the pairs of observers, both values were .81. Thus, generalizability theory, unlike classical test theory, recognizes that reliability estimates must consider the type of decision to be made with the data in hand.

## D-study Analyses

In addition to allowing the researcher to evaluate the quality of data given different the types of decisions to be made, another unique positive feature of generalizability theory is its ability to assist the researcher in making decisions about the measurement protocols themselves. As Shavelson, Webb and Burstein (1985, p. 66) explain,

> A major contribution of generalizability theory is that it allows the researcher to pinpoint the sources of measurement error (e.g., rater, occasion, or both) and increase the appropriate number of observations accordingly so that error "averages out." The researcher can estimate how many conditions of each facet are needed to obtain a certain level of generalizability.

These analyses are termed "D"-studies. Table 2 presents the estimated effects of certain variations in the measurement protocol. These results can be consulted to determine where measurement protocols are uneconomical in yielding improved reliability, and the extent to which economical variations yield improved measurement quality.

INSERT TABLE 2 ABOUT HERE.

For example, the tabled results indicate that asking each nurse observer to make three measurements yields little if any improvement in generalizability; the results suggest that one measurement would yield equally good data regarding the quality of the pairs of nurses. The results indicate that increasing the number of measured individuals readily yields substantial improvements, so that measurements involving more than 50 or 75 individuals may not be worthwhile, if assessing observer quality is the researcher's primary interest.

## Discussion

Several comments regarding the proper use of generalizability theory may be in order prior to enumerating the advantages of the theory. First, as Webb, Shavelson, Shea and Morello (1981, p. 191) note:

> Unless many levels of each [measurement design] facet are sampled, confidence intervals for the variance components may be very wide. Because estimated variance components are the basis for indexing the relative contributions of each source of error, valid interpretations depend on stable estimates.

In fact, some researchers suggest that improvements are realized by using up to as many as 10 levels per measurement protocol facet (Calkins, Erlich, Marston & Malitz, 1978).

It should also be noted that non-nested designs may usually be preferable for "G"-studies. However, Shavelson and Webb (1981, p. 134) argue that "A nested G study is sometimes useful because

it provides more degrees of freedom for some estimates of variance components." Monte Carlo studies by Smith (1980) indicate that some nested designs produce more stable variance components estimates than do non-nested designs.

Finally, it should be noted that the researcher's decisions about the population to which generalizations are to be made, i.e., whether the levels of a facet represent the full universe or a sample from that universe, affect results. Furthermore, declarations of interest in a given object of measurement also affect results. For example, in the results tabled in the present report it was determined that statements about the quality of the pairs of nurse observers were of interest. The results would have varied if the researcher had declared primary interest in making judgments about the individuals in the study.

Thus, perhaps the most important benefit from the use of generalizability theory is that it forces the researcher to make careful decisions about what measurement questions are being asked and about the population to which results are to be generalized. The theory is also valuable in that it allows the calculation of different coefficients given interest in making either norm-referenced or criterion-referenced decisions. Finally, the theory supports the analysis of "what if" questions that evaluate the value of variations in the measurement protocol and the economics of the protocol.

Rowley (1976, p. 51) has observed that

> The variance components approach... enables the researcher to pinpoint multiple sources of error, and to compute a number of different reliability

coefficients for different purposes. Unfortunately, the literature does not indicate that these methods have gained wide acceptance, at least not in practice.

However, Dyer (1979, p. 125) has predicted that "Coefficients of generalizability will become more common in the research literature as more researchers assimilate these techniques into their repertoire of skills." The preceeding enumeration of the benefits of the theory suggests that Dyer's prediction may well be realized.

## References


Best, J. W. (1981). Research in education (4th ed.). Englewood Cliffs: Prentice-Hall.

Borg, W. R., & Gall, M. D. (1983). Educational research: An introduction (4th ed.). New York: Longman.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: The American College Testing Program.

Calkins, D. S., Erlich, O., Marston, P. T., & Malitz, P. (March, 1978). An empirical investigation of the distributions of generalizability coefficients and variance estimates for an application of generalizability theory. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Cates, W. M. (1985). A practical guide to educational research. Englewood Cliffs: Prentice-Hall.

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), Handbook of multivariate psychology. Chicago: Rand-McNally, 67-128.

Crick, J. E., & Brennan, R. L. (1983). Manual for GENOVA: A GENeralized analysis Of VAriance system (ACT Technical Bulletin No. 43). Iowa City, IA: The American College Testing Program.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New

York: Wiley.

Dyer, J. R. (1979). Understanding and evaluating educational research. Reading, MS: Addison-Wesley.

Hopkins, C. D. (1980). Understanding educational research: An inquiry approach. Columbus, OH: Merrill.

McMillan, J. H., & Schumacher, S. (1984). Research in education: A conceptual introduction. Boston: Little, Brown and Company.

Mercatores, M., & Craighead, E. (1974). Effects of nonparticipant observation in teacher and pupil classroom behavior. Journal of Educational Psychology, 66, 512-519.

Mouly, G. J. (1978). Educational research: The art and science of investigation. Boston: Allyn and Bacon.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Salvia, J. A., & Mersel, C. J. (1980). Observer bias: A methodological consideration in special education research. Journal of Special Education, 14, 261-270.

Sax, G. (1980). Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, CA: Wadsworth.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1985). In M. C. Wittrock (Ed.), Handbook of Research on Teaching (3rd ed.). New York: MacMillan.

Smith, P. L. (April, 1980). Some approaches to determining the stability of estimated variance components. Paper presented at the annual meeting of the American Educational Research

Association, Boston.

Webb, N. M., Shavelson, R. J., Shea, J., & Morello, E. (1981). Generalizability of general educational development ratings of jobs in the United States. Journal of Applied Psychology, 66, 186-192.

Voors, A. W., Foster, T. A., Freuchs, R. C., Webber, L. S., & Berenson, G. S. (1976). Studies of blood pressure in children, ages 5-14 years, in a total biracial community--The Bogalusa Heart Study. Circulation, 54, 319-327.

(caption)

Figure 1

Graphic Representation of the Measurement Protocol

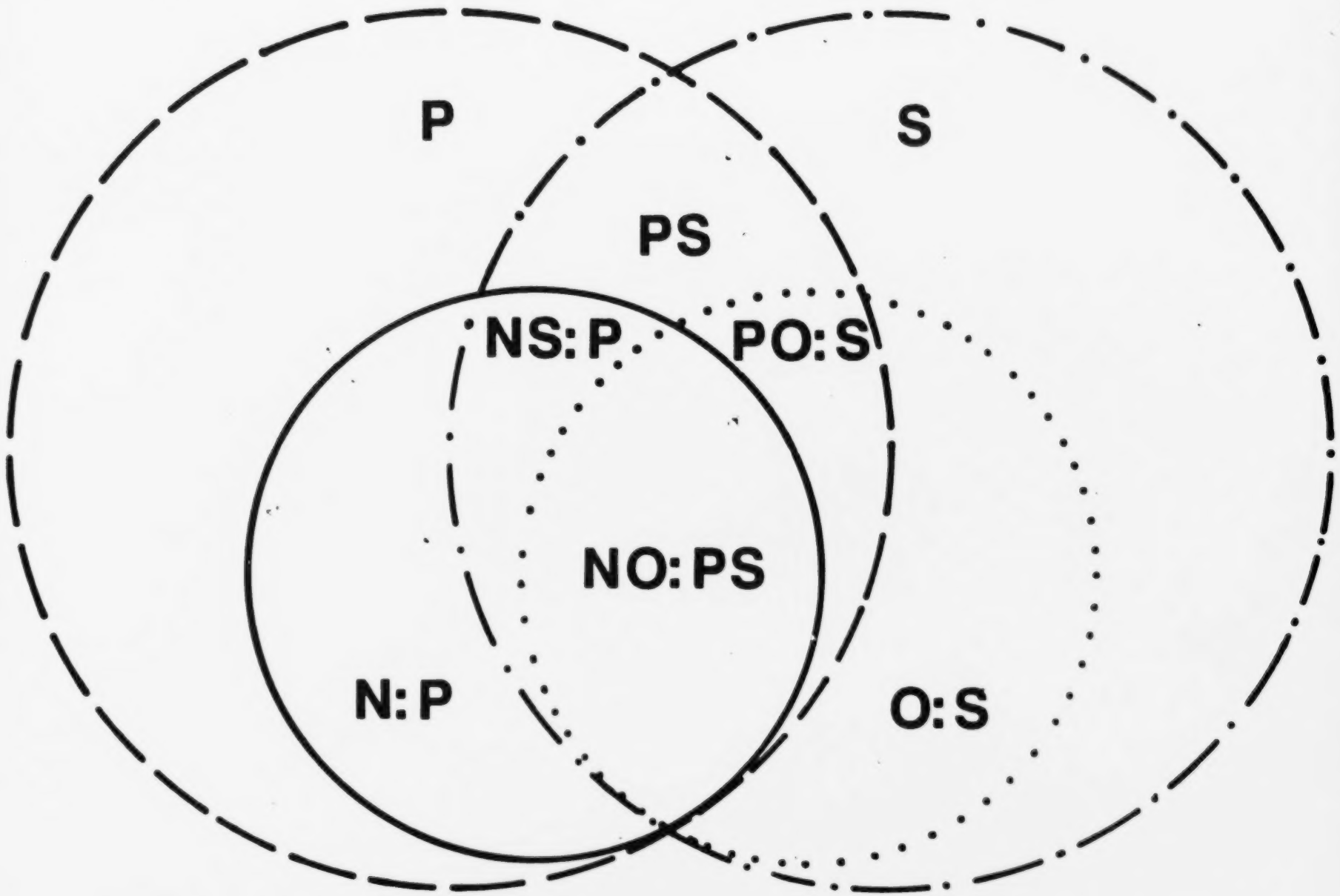
P
S
PS
NS:P
PO:S
NO:PS
N:P
O:S

Table 1
Variance Components

| Source | df | SOS for Mean Scores | SOS for Scores | Mean Squares | Variance Component |
|---|---|---|---|---|---|
| (P)airs | 5 | 6422626.3 | 11238.0 | 2247.6 | 17.69 |
| (I)ndividuals:P | 96 | 6465169.3 | 42543.0 | 443.2 | 72.82 |
| (S)ets | 1 | 6411502.1 | 113.9 | 113.9 | 0.20 |
| (O)bservations:S | 4 | 6411612.5 | 110.4 | 27.6 | 0.21 |
| PS | 5 | 6422934.7 | 194.6 | 38.9 | 0.16 |
| PO:S | 20 | 6423139.3 | 94.1 | 4.7 | (0.00) |
| IS:P | 96 | 6468442.7 | 2964.9 | 30.9 | 8.21 |
| IO:PS | 384 | 6471048.0 | 2400.8 | 6.2 | 6.25 |
| Mean | | 6411388.2 | | | |
| Total | 611 | | 56659.8 | | |

Note. P and S were considered "fixed" effects, while I and O were considered "random" effects.

Table 2
"D"-study Results

| Source | P | I | S | O | Generalizability | Phi |
|---|---|---|---|---|---|---|
| Universe Size | 6 | Inf | 2 | Inf | Coefficient | Coefficient |
| Protocol Variation | 6 | 17 | 2 | 1 | .798 | .795 |
| | 6 | 17 | 2 | 2 | .802 | .800 |
| | 6 | 17 | 2 | 3 | .803 | .802 |
| | 6 | 17 | 2 | 4 | .803 | .802 |
| | 6 | 17 | 2 | 5 | .804 | .803 |
| | 6 | 50 | 2 | 3 | .923 | .921 |
| | 6 | 75 | 2 | 3 | .947 | .946 |
| | 6 | 100 | 2 | 3 | .960 | .958 |
| | 6 | 200 | 2 | 3 | .980 | .978 |